

Sequential motif profile of natural visibility graphs

Jacopo Iacovacci, Lucas Lacasa*

School of Mathematical Sciences, Queen Mary University of London, Mile End Road, E14NS London (UK)

The concept of sequential visibility graph motifs -subgraphs appearing with characteristic frequencies in the visibility graphs associated to time series- has been advanced recently along with a theoretical framework to compute analytically the motif profiles associated to Horizontal Visibility Graphs (HVGs). Here we develop a theory to compute the profile of sequential visibility graph motifs in the context of Natural Visibility Graphs (VGs). This theory gives exact results for deterministic aperiodic processes with a smooth invariant density or stochastic processes that fulfil the Markov property and have a continuous marginal distribution. The framework also allows for a linear time numerical estimation in the case of empirical time series. A comparison between the HVG and the VG case (including evaluation of their robustness for short series polluted with measurement noise) is also presented.

PACS numbers:

I. INTRODUCTION

In recent years different methods [1–5] have been proposed to map the structure and underlying dynamics of a given time series into an associated graph representation, with the aims of exploiting the modern tools of network science [6–8] in the traditional task of time series analysis [9, 10], thereby building a bridge between the two fields. In this context, visibility graphs have been proposed [11, 12] as a tool to extract a graph from the relative positions of an ordered series, from which several graph features can be extracted and used for description and classification problems. Very recently we have advanced the concept of sequential visibility graph motifs [13], building on the idea of network motifs [14, 15] to explore the decomposition of visibility graphs into sequentially restricted subgraphs. These motifs induce a graph-theoretical symbolization of a given time series into a sequence of subgraphs. We have shown that the marginal distribution of the motif sequence -the so-called motif profile- is an informative feature to describe different types of dynamics and is useful in the task of classifying empirical time series. For large classes of dynamical systems, we were able to develop a theory to analytically compute the frequency of each motif when these are extracted from a so-called Horizontal Visibility Graph (HVG), this being a modified and simpler version of the original (natural) visibility graph (VG) which has often shown analytical tractability [16–19]. As a matter of fact, in the case of VGs to obtain analytical insight has shown to be a challenging task, and besides few exception [20] most of the works that make use of this statistic are computational. In this paper we bridge this gap and advance a theory to analytically compute the complete motif profile in the natural case (VG motifs). We focus on motifs of size $n = 4$ as this was shown to be the simplest case which gives nontrivial results [13]. We validate this theory by deriving explicit motif profiles for several classes of dynamics which we show to be in good agreement with numerical simulations. We also study the robustness of this feature when the time series is short and polluted with measurement noise, and compare its performance with the case of HVG motifs.

The rest of the paper is as follows: after recalling the definitions of natural and horizontal visibility graphs, in Section II we present the concept and main properties of sequential visibility graph motifs, as well as recalling the theoretical framework where the motif profile from the horizontal version was derived. In Section III we focus on natural visibility and develop the theory to compute analytically the motif profile associated to processes where the dynamics are either bounded or unbounded. We test this theory by assessing the predictions for different dynamical systems, and we also show that white noise with different marginals can be distinguished using the natural version instead of the horizontal one. In Section IV we show that the visibility graph motif profile is a robust feature in the sense of (i) having a fast convergence to asymptotic values for short series size N and (ii) being robust against contamination with measurement noise (white and colored). In Section V we conclude.

*Electronic address: j.iacovacci@qmul.ac.uk; l.lacasa@qmul.ac.uk

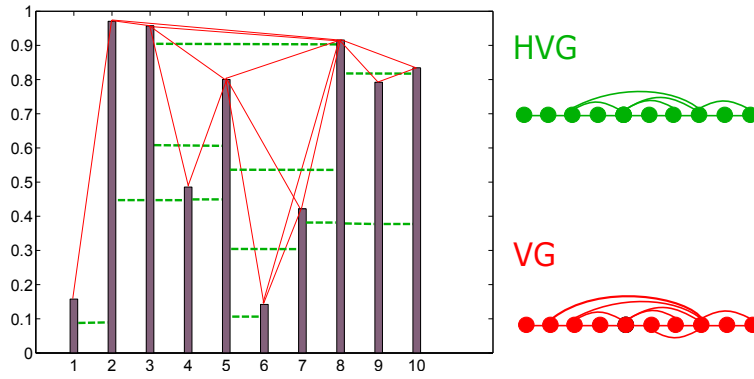


FIG. 1: (Color online) Two visibility algorithms, Natural Visibility (red) and Horizontal Visibility (green) applied to a time series of 10 data (bars); the corresponding Visibility Graph (VG) and Horizontal Visibility Graph (HVG) are shown on the right: each datum in the series corresponds to a node in the graph and two nodes are connected if their corresponding data heights show respectively natural visibility or horizontal visibility (see the text).

II. VISIBILITY GRAPHS AND SEQUENTIAL MOTIFS

Let $\mathbf{x} = \{x_1, \dots, x_N\}$ be a real-valued time series of N data. The natural visibility graph (VG) [11] extracted from the series is the graph $\mathcal{G} = \{V, E\}$ where each datum x_i in the series is associated to a node v_i (thus $|V| = N$ and V is a totally ordered set) and an edge $e_{i,j} \in E$ between node i and node j exists if $x_k < x_i + \frac{k-i}{j-i}[x_j - x_i]$ for each k such that $i < k < j$. This is called the visibility criterion, which for the natural version is indeed a convexity criterion. Analogously, the horizontal visibility graph (HVG) [12] extracted from the series is the graph $\mathcal{G}^h = \{V, E^h\}$, with the same vertex set than \mathcal{G} and a smaller edge set E^h , where an edge $e_{i,j} \in E$ exists between nodes i and j if $x_k < \max(x_i, x_j)$, $\forall k : i < k < j$. This visibility criterion is in turn an ordering one. It was indeed shown that the HVG associated to a time series is invariant under monotonic transformations in the series [21], thus the HVG is an order statistic of the series. \mathcal{G}^h is indeed a non crossing graph [22] which by construction is always also a sub-graph of \mathcal{G} (although \mathcal{G} is not in general planar). Both VG and HVG are connected graphs with a trivial Hamiltonian path given by the sequence of vertices (v_1, v_2, \dots, v_N) . An illustration of how to construct a VG and HVG from a given time series is shown in figure 1.

The set of sequential VG motifs of size n , $n \in [2, 3, \dots, N]$ is defined as the set of all the M_n possible sub-graphs with n consecutive vertices along the Hamiltonian path of a VG (similarly, the set of HVG motifs of size n is the set of all the M_n^h admissible sub-graphs with n consecutive vertices along the Hamiltonian path of a HVG). Accordingly, sequential VG motifs are also visibility graphs. For $n = 4$, there are in principle a total of $M_4 = 8$ possible motifs (see table II for an enumeration), although as we will show below the number of admissible ones is just 6. Given a VG \mathcal{G} , its sequential motifs can be detected using a sliding window of size n which slides along the Hamiltonian path of the graph with $N - n$ consecutive overlapping steps. At each step a particular motif is detected inside the window. We can accordingly estimate Φ_m , the frequency of appearance of a certain motif m , and define the n -motif profile $Z^n = (\Phi_1, \dots, \Phi_{M_n})$. The process of extracting a VG/HVG and its sequential visibility motif set is illustrated in Figure 2 (the concept is analogous for HVG, although the set of admissible motifs is different in both cases). Note that since Z^n can be understood as a discrete probability distribution and is therefore a vector with unit norm (we use the ℓ_1 norm here) $\sum_{m=1}^{M_n} \Phi_m = 1$, the number of degrees of freedom of Z^n is $M_n - 1$ (again, as we will see below, in the case considered here it is even more reduced as the number of admissible motifs will be less than 8).

In a recent work [13] we introduced the concept of sequential HVG motifs and advanced a theory to compute in an exact way Z^4 in the case of the HVG. It was shown that the $n = 4$ motif statistic was useful to discriminate across different types of dynamics. The case of uncorrelated noise was shown to yield a universal motif profile, independent of the marginal distribution of the i.i.d. process and this enabled the definition of randomness test. We also found for some deterministic dynamics some *forbidden motifs*, which represented a persistent characteristic to test the randomness of a process (note that if a motif of size n doesn't occur, then also all the motifs of size $n' > n$ which incorporate that motif won't occur either). Since VG and HVG n motif profiles are a temporally constrained feature (they are evaluated along n consecutive nodes on the Hamiltonian path) its extraction can be seen as a process of dynamic symbolization. Under this perspective, the relation between HVG motifs of size $n = 4$ and the so-called *ordinal patterns* (OPs) [23] was acknowledged in [13]. n -OPs are symbols extracted from a time series representing the

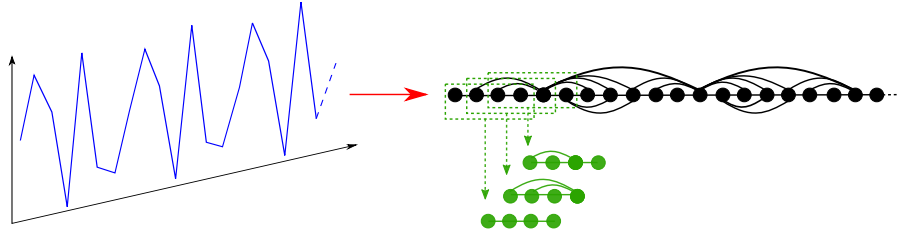


FIG. 2: Schematic of visibility graph motif detection. A time series is converted into a visibility graph according to the visibility criterion (red arrow). A window of size $n = 4$ slides along the Hamiltonian path of the VG graph and detects at each step a different VG motif.

possible ranking output of n consecutive data and are extracted from a specific time series by comparing the value of all the set of n consecutive data along the series [23–25]. It was not unexpected to find a link between HVG n -motifs and n -OPs as HVG is known to be an order statistic, much as OPs. Indeed, in the particular case of a time series for which data don't repeat $P(x_t = x_{t+1}) \simeq 0$ there exists a mapping between each appearing HVG motif and a specific set of ordinal patterns [13]; in this scenario the forbidden motifs selected by the horizontal visibility are, in general, set of the so called *forbidden ordinal patterns* [26]. Of course both VG and HVG motifs analysis can be applied without requiring any further assumption to time series taking values from finite sets (namely when $P(x_t = x_{t+1}) \neq 0$), while the ordinal patterns approach -based uniquely on the ranking statistics- require further assumptions in that case. Here we focus in the *natural* version of the algorithm and explore VG motifs instead. As VGs are not invariant under monotonic transformations in the series [21], in general they depend on the marginal probability distribution of the time series and are not an order statistic. Accordingly, there is no obvious correspondence between n -OPs and VG n -motifs and both approaches in principle represent two independent symbolization methods that encode temporal information in a different way. In what follows we recall the theoretical framework for HVG motifs and in the next section we extend this theory to deal with VG motifs.

A. Theory for the HVG motif profile

Label	HVG motif type	Inequality set
1	••••	$\{\forall(x_l, x_{l+1}), x_{l+2} < x_{l+1}, x_{l+3} < x_{l+2}\} \cup \{\forall(x_l, x_{l+3}), x_{l+1} > x_l, x_{l+2} > x_{l+1}\}$
2	••••	$\{\forall x_l, x_{l+1} < x_l, x_{l+2} = x_{l+1}, x_{l+3} > x_{l+2}\}$
3	••••	$\{\forall x_l, x_{l+1} < x_l, x_{l+1} < x_{l+2} < x_l, x_{l+3} < x_{l+2}\} \cup \{\forall(x_l, x_{l+3}), x_{l+1} < x_l, x_{l+2} > x_l\}$
4	••••	$\{\forall x_l, x_{l+1} > x_l, x_{l+2} < x_{l+1}, x_{l+3} > x_{l+2}\} \cup \{\forall x_l, x_{l+1} < x_l, x_{l+2} < x_{l+1}, x_{l+2} < x_{l+3} < x_{l+1}\}$
5	••••	$\{\forall x_l, x_{l+1} < x_l, x_{l+1} < x_{l+2} < x_l, x_{l+3} > x_{l+2}\}$
6	••••	$\{\forall x_l, x_{l+1} < x_l, x_{l+2} < x_{l+1}, x_{l+3} > x_{l+1}\}$

TABLE I: The set of size-4 HVG motifs are defined according to a set of relations between 4 arbitrary consecutive data $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$, $l \in [1, N - 3]$ in the time series.

Consider a generic dynamical process $\mathcal{H} : \mathbb{R} \rightarrow \mathbb{R}$ with a smooth invariant measure $f(x)$, $x \in (a, b)$ that fulfils the Markov property $f(x_l | x_{l-1}, x_{l-2}, \dots) = f(x_l | x_{l-1})$, where $f(x_l | x_{l-1})$ is the transition probability distribution. It was shown in [13] that, given four arbitrary consecutive data $x_l \dots x_{l+3}$ the motif profile $Z^4 = (\Phi_1^4, \dots, \Phi_6^4)$ could then be computed as

$$\Phi_m^4 = \int f(x_l) dx_l \int f(x_{l+1} | x_l) dx_{l+1} \dots \int f(x_{l+n-1} | x_{l+n-2}) dx_{l+n-1} \quad (1)$$

where the range of each integral was implicitly given by the inequality set reported in table I. This inequality set encoded the relative location of data within a particular motif, and indeed highlights the order statistic nature of this measure. In those experimental cases where there is no access to $f(x_t | x_{t-1})$, this approach allowed in turn for an numerical estimation of HVG motif profile with linear time complexity $O(N)$.

III. THEORY FOR SEQUENTIAL VG MOTIFS

Let us consider again a time-discrete (deterministic or stochastic) dynamical process $x_{t+1} = \mathcal{H}(x_t, \xi)$ that fulfils the Markov property: $\forall l f(x_l|x_{l-1}, x_{l-2}, \dots) = f(x_l|x_{l-1})$, where $f(x_l|x_{l-1})$ is the transition probability distribution and $x \in (a, b)$. For deterministic processes $f(x_l|x_{l-1}) = \delta(x_l - \mathcal{H}(x_{l-1}))$ where $\delta(x)$ is the Dirac-delta distribution where $\delta(x)$ is the Dirac-delta distribution:

$$\int_p^q \delta(x - y) dx = \begin{cases} 1 & y \in [p, q] \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

and $f(x)$ is a smooth invariant measure of the process $\mathcal{H}(x)$, whereas for stochastic processes $f(x)$ is simply the underlying probability density, i.e. the marginal distribution of the process. Our theory addresses the motif profile Z^4 , in what follows we split this analysis in two cases, depend whether x is bounded or unbounded. In both cases, each probability Φ_m^4 is computed formally using concatenated integrals which are formally equivalent to eq.1, where the ranges of each integral are given according to the convexity criteria defining the visibility rule, as opposed to the HVG case where these were simply ordering criteria. Whereas in the case of unbounded variables the inequality set will only take into account the visibility criteria within the motifs, in the case of bounded variables the additional restriction of variables needing to be bounded adds a layer of complexity as we will see. From now on, let $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$ be four arbitrary consecutive data ($l \in [1, N - 3]$).

A. Unbounded variable $x \in (-\infty, \infty)$

Label	VG motif type	Inequality set
1		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, x_{l+3} < 2x_{l+2} - x_{l+1}\}$
2		$\{\emptyset\}$
3		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, x_{l+3} < \frac{3}{2}x_{l+2} - \frac{1}{2}x_l\}$
4		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, 2x_{l+2} - x_{l+1} < x_{l+3} < 3x_{l+1} - 2x_l\}$
5		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, \frac{3}{2}x_{l+2} - \frac{1}{2}x_l < x_{l+3} < 2x_{l+2} - x_{l+1}\}$
6		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, x_{l+3} > 3x_{l+1} - 2x_l\}$
7		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, x_{l+3} > 2x_{l+2} - x_{l+1}\}$
8		$\{\emptyset\}$

TABLE II: The set of size-4 VG motifs are defined according to a set of relations between 4 consecutive data $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$, $l \in [1, N - 3]$ in the time series.

In the case of unbounded variables, it is easy to prove that in general

$$\Phi_m^n = \int_{\mathbb{R}} f(x_l) dx_l \int_{d_1^m(x_l)}^{c_1^m(x_l)} f(x_{l+1}|x_l) dx_{l+1} \dots \int_{d_{n-1}^m(x_l, \dots, x_{l+n-2})}^{c_{n-1}^m(x_l, \dots, x_{l+n-2})} f(x_{l+n-1}|x_{l+n-2}) dx_{l+n-1} \quad (3)$$

where $\{c_i^m(\cdot)\}_{i=1, \dots, n-1}$ and $\{d_i^m(\cdot)\}_{i=1, \dots, n-1}$ are the set of functions which specify respectively the upper bound condition and the lower bound condition for the i -th integral. As advanced, these conditions are directly related to the visibility criterion (which in the VG case is a convexity relation) and a summary of those are explicitly reported in table II (the equivalence between each motif and its associated inequality set can be proved rigorously although we omit here this proof as it is quite trivial). First, note that for the motifs 2 and 8 the inequality set is empty. This means that these motifs are actually not admissible under the VG algorithm. In the case of motif number 2, note that this motif was an admissible one for HVGs associated to integer-valued series, where the probability of finding equal consecutive data is finite. For VG, it is easy to prove that if the bounding nodes share an edge, then either the left edge or the right edge shares an edge with an inner node. Thus motif 2 is not a VG. Similarly, it is easy to prove that if a time series gives rise to a motif of type 8, then an edge would necessarily appear between the two bounding nodes, reducing this to type 7. Accordingly, the number of admissible motifs is not 8 but 6, and thus the effective number of degrees of freedom associated to $n = 4$ VG motifs is $M_n - 2 - 1 = 5$.

To better understand the application of the inequality set in the case of unbounded variables, consider a white Gaussian process ($x_i \in (-\infty, \infty)$) with

$$f(x_i) = \frac{\exp(-x_i^2/2)}{\sqrt{2\pi}} \quad \text{and} \quad f(x_{i+1}|x_i) = f(x_{i+1})$$

the probability of appearance of motif 1 in Table II can be written explicitly as

$$\Phi_1^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{-\infty}^{2x_1-x_0} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{-\infty}^{2x_2-x_1} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3. \quad (4)$$

and the integral can be easily evaluated up to arbitrary precision to obtain $\Phi_1^4 \simeq 0.13386$.

B. Bounded variables $x \in [a, b]$; $a, b < \infty$

In the case where $x \in (a, b)$ where the bounds $a, b < \infty$ are finite, these restrictions in turn induce further conditions on the lower and upper bounds of the integrals in Eq.3 which have the effect of splitting the overall integral in a sum of different integrals. For illustrative purposes we start by considering a particular example. Consider a series of i.i.d. uniform random variables $x_i \sim \mathcal{U}[a, b]$ with $f(x) = (b-a)^{-1}$ and $f(x_{i+1}|x_i) = f(x_{i+1})$, and let us consider again Φ_1^4 . According to table II, in principle the conditions for the first two variables x_0, x_1 are $\forall(x_0, x_1) \in [a, b]$; for the third variable x_2 the lower bound condition becomes $x_2 > a$ but the upper bound condition will depend on the function $2x_1 - x_0$ which can take values in $[2a - b, 2b - a]$ and thus we need to consider three different cases:

$$\left\{ \begin{array}{l} 2x_1 - x_0 > b \implies x_1 \in (\frac{(x_0+b)}{2}, b], \quad x_2 \in [a, b] \\ a < 2x_1 - x_0 < b \implies x_1 \in (\frac{(x_0+a)}{2}, \frac{(x_0+b)}{2}], \quad x_2 \in [a, 2x_1 - x_0] \\ 2x_1 - x_0 < a \quad \quad \quad \{\emptyset\} \end{array} \right. \quad (5)$$

where the last case doesn't contribute (as $x_2 > a$ is always fulfilled). Similarly for each admissible choice of the variable x_2 the bound conditions for x_3 will produce an additional split:

$$\left\{ \begin{array}{l} 2x_2 - x_1 > b \implies x_2 \in (\frac{(x_1+b)}{2}, b], \quad x_3 \in [a, b] \\ a < 2x_2 - x_1 < b \implies x_2 \in (\frac{(x_1+a)}{2}, \frac{(x_1+b)}{2}], \quad x_3 \in [a, 2x_2 - x_1] \\ 2x_2 - x_1 < a \quad \quad \quad \{\emptyset\} \end{array} \right. \quad (6)$$

After a bit of algebra one finds

$$\Phi_1^4 = \frac{1}{(b-a)^4} \left[\int_a^b dx_0 \int_{\frac{(x_0+b)}{2}}^b dx_1 \int_{\frac{(x_1+b)}{2}}^b dx_2 \int_a^b dx_3 + \right. \quad (7)$$

$$+ \int_a^b dx_0 \int_{\frac{(x_0+b)}{2}}^b dx_1 \int_{\frac{(x_1+a)}{2}}^{\frac{(x_1+b)}{2}} dx_2 \int_a^{2x_2-x_1} dx_3 + \quad (8)$$

$$+ \int_a^b dx_0 \int_{\frac{(2x_0+b)}{3}}^{\frac{(x_0+b)}{2}} dx_1 \int_{\frac{(x_1+b)}{2}}^{2x_1-x_0} dx_2 \int_a^b dx_3 + \quad (9)$$

$$+ \int_a^b dx_0 \int_{\frac{(2x_0+a)}{3}}^{\frac{(2x_0+b)}{3}} dx_1 \int_{\frac{(2x_1+a)}{2}}^{2x_1-x_0} dx_2 \int_a^{2x_2-x_1} dx_3 + \quad (10)$$

$$\left. + \int_a^b dx_0 \int_{\frac{(2x_0+b)}{3}}^{\frac{(x_0+b)}{2}} dx_1 \int_{\frac{(x_1+a)}{2}}^{\frac{(x_1+b)}{2}} dx_2 \int_a^{2x_2-x_1} dx_3 \right] = \frac{5}{36} \simeq 0.1389. \quad (11)$$

Two comments are in order. First, note that this result is different from the value for Φ_1^4 for Gaussian white noise, that is, the results for white noise seems to be dependent on the marginal distribution of the noise. This lack of invariance was expected as VG is not an order statistic, and differs from the phenomenology found for HVG, where results for white noise were universal (independent from the marginal distribution). This evidence will be confirmed in the next sections. Second, for uniform white noise this result appears independent from the values a and b , suggesting that results could depend on the marginal but be independent upon rescaling of the support (a, b) . This is however not the case in general, and this independence seems to be unique for uniformly distributed white noise (results not shown). After a bit of algebra, we are able to translate the visibility and bounded variable restrictions inside each motif into another set of inequalities, which we have reported in table III. In what follows we make use of this theory to compute the theoretical VG motif profile for several dynamical processes.









Motif label	Motif type	Inequality set
1		$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, b], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
2		$\{\emptyset\}$
3		$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_l+2b)}{3}, b], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_l+2a)}{3}, \frac{(x_l+2b)}{3}], x_{l+3} \in [a, \frac{3x_{l+2}-x_l}{2}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [2x_{l+1} - x_l, b], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_l+2b)}{3}, b], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_l+2b)}{2}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(x_l+2a)}{3}, \frac{(x_l+2b)}{3}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$
4		$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [2x_{l+2} - x_{l+1}, 3x_{l+1} - 2x_l]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$
5		$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+2a)}{3}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(2x_l+a)}{3}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_l+2a)}{3}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_l+2a)}{3}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$ $\{x_l \in [a, b], x_{l+1} \in [a, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, \frac{(x_l+2b)}{3}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, b]\}$
6		$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [a, 2x_{l+1} - x_l], x_{l+3} \in [3x_{l+1} - 2x_l, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [a, 2x_{l+1} - x_l], x_{l+3} \in [a, b]\}$
7		$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$ $\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$
8		$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, \frac{(x_{l+1}+b)}{2}]\}$ $\{\emptyset\}$

TABLE III: Sets of inequalities between 4 consecutive data $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$, $l \in [1, N-3]$ in a time series of length N which define the VG motifs of size 4 in the case of bounded variables $x_i \in [a, b]$.

C. VG motif profiles of different families of dynamical systems

1. Fully chaotic logistic map

We start by considering the fully chaotic logistic map $\mathcal{H}(x) = 4x(1-x)$, $x \in [0, 1]$, with invariant density $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. As this process is deterministic, it fulfils a trivial Markov property such that $f(x_2|x_1) = \delta(x_2 - \mathcal{H}(x_1))$. The HVG motif profile for this process was computed exactly in [13], here we compute the VG motif profile. Before proceeding to compute each probability contribution, it is important to highlight a subtle point. Since for this process $x \in [0, 1]$ is bounded, in principle one should use the inequality set depicted for bounded variables in table III. However, in this particular case it is actually not necessary to explicitly consider the restriction $x \in [0, 1]$. As we will see in a moment, this is already taken into account implicitly in the computation of each integral and therefore one can use the (simpler) inequality set for unbounded variables given in table II.

We start by computing Φ_1^4 :

$$\Phi_1^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1-x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{2x_2-x_1} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) < 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.1743, 0.25]$. Note at this point that the latter conditions are also satisfied in other ranges, but we only consider those ranges that belong to $[0, 1]$, and this is indeed the reason why we don't need to use in this case the inequality set for bounded variables. We thus have

$$\Phi_1^4 \simeq \frac{1}{\pi} B_{[0.1743, 0.25]} \left(\frac{1}{2}, \frac{1}{2} \right) \simeq 0.0591$$

As $\Phi_2^4 = 0$ by construction, we proceed by calculating Φ_3^4 :

$$\Phi_3^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{2x_1-x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{\frac{3}{2}x_2-\frac{1}{2}x_0} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) < \frac{3}{2}\mathcal{H}^2(x_0) - \frac{1}{2}\mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.0522, 0.1743] \cup [0.75, 0.929]$. Therefore:

$$\Phi_3^4 \simeq \frac{1}{\pi} (B_{[0.0522, 0.1743]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.75, 0.929]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.289.$$

Similarly for Φ_4^4 we have

$$\Phi_4^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1-x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{2x_2-x_1}^{3x_1-2x_0} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$2\mathcal{H}^2(x_0) - \mathcal{H}(x_0) < \mathcal{H}^3(x_0) < 3\mathcal{H}^2(x_0) - 2x_0$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.25, 0.75]$, and therefore

$$\Phi_4^4 \simeq \frac{1}{\pi} B_{[0.25, 0.75]} \left(\frac{1}{2}, \frac{1}{2} \right) \simeq 0.3333$$

For Φ_5^4 we have

$$\Phi_5^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{2x_1-x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{\frac{3}{2}x_2-\frac{1}{2}x_0}^{2x_2-x_1} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\frac{3}{2}\mathcal{H}^2(x_0) - \frac{1}{2}x_0 < \mathcal{H}^3(x_0) < 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.927, 0.954]$, and thus

$$\Phi_5^4 \simeq \frac{1}{\pi} (B_{[0.04568, 0.05224]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.9239, 0.9543]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.0439$$

For Φ_6^4 we have

$$\Phi_6^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1-x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{3x_1-2x_0}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) > 3\mathcal{H}(x_0) - 2x_0$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are never satisfied and thus

$$\Phi_6^4 = 0.$$

For Φ_7^4 we have

$$\Phi_7^4 = \int_0^1 f(x_0)dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0))dx_1 \int_{2x_1-x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0))dx_2 \int_{2x_2-x_1}^1 \delta(x_3 - \mathcal{H}^3(x_0))dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) > 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0, 0.0457] \cup [0.9544, 1]$, and thus

$$\Phi_7^4 \simeq \frac{1}{\pi} (B_{[0, 0.046]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.95, 1]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.2741$$

Finally, by construction $\Phi_8^4 = 0$. Altogether, we find the VG motif profile of a fully chaotic logistic map

$$\mathbf{Z}^4 = (0.0591, 0, 0.289, 0.3333, 0.0439, 0, 0.2741, 0) \quad (12)$$

Note that while the result is in this case an approximation, our theory allows for numerical estimates with arbitrary precision (the result is not exact because the location of fixed points of the map is only approximate, although this approximation is arbitrarily close to the true values).

2. Uniform white noise

For white uniform noise $x(t) = \xi$, $\xi \sim U[a, b]$ we have a probability density $f(x)$ and transition probability $f(x_2|x_1)$ given by

$$f(x) = \frac{1}{b-a}, \text{ and } f(x_2|x_1) = f(x_2) \quad (13)$$

In this case the computations are more cumbersome since we need to make use of the inequality set for bounded variables described in table III. The m th component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \sum_s \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{d_{s1}^m}^{c_{s1}^m} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{d_{s2}^m}^{c_{s2}^m} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{d_{s3}^m}^{c_{s3}^m} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3 \quad (14)$$

where the sum runs over all the set s of conditions d_{si}^m, c_{si}^m which contribute to evaluate the probability for motif m in Table III. All the integrals can nonetheless be solved analytically in closed form and give the following motif profile

$$\mathbf{Z}^4 = \left(\frac{5}{36}, 0, \frac{31}{108}, \frac{31}{108}, \frac{2}{27}, \frac{2}{27}, \frac{5}{36}, 0 \right) \quad (15)$$

Several comments are in order. First, this profile is different from the one found for the chaotic logistic map. Second, this motif profile turns to be independent from the bounding values a and b (where $x \in [a, b]$), meaning that white, uniform noise has a VG motif profile which is invariant under various transformations in the original distribution of the time series. This is not a trivial property and is indeed a peculiarity of the uniform distribution, in other words the VG motif profile of white noise extracted from a bounded distribution *generally* depends on the bounds of the distribution.

3. Gaussian white noise

For standard white Gaussian noise $x(t) = \xi$, $\xi \sim \mathcal{N}(0, 1)$ the probability density $f(x)$ and transition probability $f(x_2|x_1)$ are given by

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}, \text{ and } f(x_2|x_1) = f(x_2) \quad (16)$$

The m component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{d_2^m}^{c_2^m} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{d_3^m}^{c_3^m} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3 \quad (17)$$

where (d_i^m, c_i^m) are now the top and bottom conditions for the variable x_{l+i} in motif m reported in Table for unbounded variables II. The integrals can be evaluated numerically up to arbitrary precision and they give the following results

$$\mathbf{Z}^4 = (0.13386, 0, 0.2850, 0.2850, 0.0811, 0.0811, 0.13386, 0) \quad (18)$$

At odds with what happens for HVG motifs [13], this result is different from the benchmark result for uniformly distributed white noise, thus there is not a universal VG motif profile for white noise as previously anticipated.

4. Gaussian red noise

Gaussian colored (red) noise with exponentially decaying correlations can be simulated using an $AR(1)$ process:

$$x_t = rx_{t-1} + \xi \quad (19)$$

where $\xi \sim \mathcal{N}(0, 1)$ is Gaussian white, and $0 < r < 1$ is a parameter that tunes the correlation. The auto-correlation function $C(t)$ decays exponentially $C(t) = e^{-t/\tau}$, where the characteristic time $\tau = 1/\ln(r)$. This model is Markovian and stationary, with a probability density $f(x)$ and transition probability $f(x_2|x_1)$ given by

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}, \text{ and } f(x_2|x_1) = \frac{\exp[-(x_2 - rx_1)^2/(2(1-r^2))]}{\sqrt{2\pi(1-r^2)}} \quad (20)$$

The m component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{(x_1 - rx_0)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_1 \int_{d_2^m}^{c_2^m} \frac{e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_2 \int_{d_3^m}^{c_3^m} \frac{e^{-\frac{(x_3 - rx_2)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_3 \quad (21)$$

where, again, (d_i^m, c_i^m) are the top and bottom conditions for the variable x_{l+i} in motif m reported in table Table II. Once set the parameter r the profile can be evaluated numerically up to arbitrary precision; here we give the profile for three possible values $r = 1/4, 1/2$ and $3/4$

$$\begin{aligned} \mathbf{Z}_{r=\frac{1}{4}}^4 &= (0.14713, 0, 0.27028, 0.27028, 0.08259, 0.08259, 0.14713, 0) \\ \mathbf{Z}_{r=\frac{1}{2}}^4 &= (0.15731, 0, 0.2595, 0.2595, 0.08316, 0.08316, 0.15731, 0) \\ \mathbf{Z}_{r=\frac{3}{4}}^4 &= (0.16410, 0, 0.25258, 0.25258, 0.08332, 0.08332, 0.16410, 0) \end{aligned} \quad (22)$$

In all these examples, theoretical results are in very good agreement with results obtained with numerical simulations reported in Figure 3.

D. Noise characterization

Differently from the HVG motifs, VG motifs statistics does not depend uniquely on the ranking statistics of the data and therefore the VG motif profile could be able in principle to discriminate white noises with different marginals.

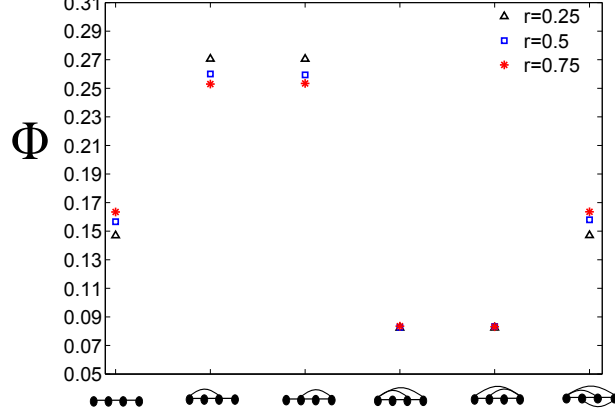


FIG. 3: Average (50 realizations) frequency of appearance Φ of VG motifs extracted from the AR(1) processes described by eq.19, for different values of the correlation coefficient $r = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}]$. Error bars are contained in the symbols; results are in very good agreement with the theoretical expected value.

In the latter sections we have been able to distinguish between Gaussian and uniform white noise. In Figure 4 we summarize the motif frequencies Φ_m^4 of VG motifs forming Z^4 , extracted from i.i.d. series with different marginals:

$$\begin{cases} \text{Uniform} \rightarrow x_i \in [0, 1]; & f(x_i) \sim 1 \\ \text{Gaussian} \rightarrow x_i \in (-\infty, \infty); & f(x_i) \sim \frac{\exp(-x_i^2/2)}{\sqrt{2\pi}} \\ \text{Power-law} \rightarrow x_i \in [1, \infty); & f(x_i) \sim x_i^{-k}, \quad k = 2.5 \\ \text{Exponential} \rightarrow x_i \in [0, \infty); & f(x_i) \sim \exp(-kx_i), \quad k = 2.5 \end{cases} \quad (23)$$

In every case we extract series of 10^5 data. The universal profile obtained for HVG is also plotted for comparison. As expected, motif profiles are different for different marginals. Motifs which are symmetric to each other (3 and 4, 5 and 6) occur with equal probabilities, something that doesn't occur when the series is chaotic (Eq. 12).

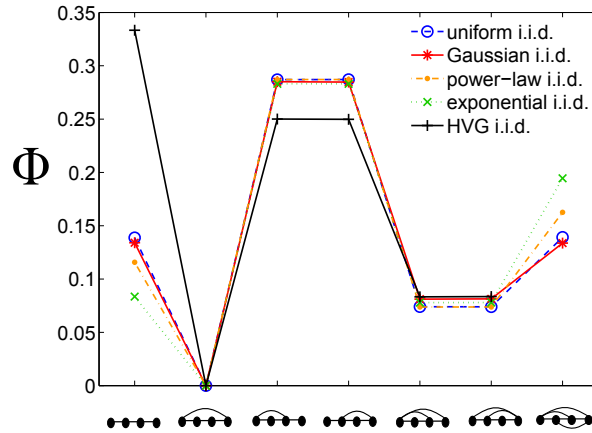


FIG. 4: Average frequency of appearance Φ of VG and HVG motifs extracted from i.i.d. series of $N = 10^5$ with different marginal distributions; VG motifs are not related with the ranking statistics of the data and able to discriminate the different types of noise.

E. Summary

According to the values obtained for the components of Z^4 , one can extract some heuristic conclusions:

- Φ_1, Φ_7 encode information on the marginal distribution of the process as well as its autocorrelation structure.
- Φ_2 is null as this motif is not a VG. This is at odds with the HVG case, where this is an admissible motif provided the probability of finding consecutive equal data in the series is finite (e.g. for discrete-valued series).
- The motifs associated to the pairs (Φ_3, Φ_4) (Φ_5, Φ_6) have chiral symmetry. In other words, the motifs associated to Φ_3 and Φ_4 are isomorphic, the correct permutation being $1 - 2 - 3 - 4 \rightarrow 4 - 3 - 2 - 1$ (the same holds for Φ_5 and Φ_6). Accordingly, for any process which is statistically time reversible, we expect these probabilities to be equal. Reversible processes include linear stochastic processes (and both white and red noise belong to this family), while non-invertible chaotic processes are usually time irreversible (the fully chaotic logistic map is an example). Time irreversibility of the process is therefore encoded in these terms.
- $\Phi_8 = 0$ as this is not a VG and therefore does not appear (not admissible).

IV. ROBUSTNESS: A COMPARISON BETWEEN VG AND HVG MOTIF PROFILES

When dealing with empirical time series, the practitioner usually faces two different but complementary challenges, namely (i) the size of the series and (ii) the possible sources of measurement noise. The first challenge can be a problem when the statistics to be extracted from the series are strongly affected by finite-size effects, whereas for the second one needs to evaluate the robustness of those statistics against noise contamination. For a statistic or feature extracted from a time series to be not just informative but useful one usually requires that statistic or feature to be robust against both problems: it needs to have fast finite-size convergence speed and to be robust against reasonably large amounts of additive noise.

In [13] it has been already shown that the HVG motif profile has good convergence properties respect to the series size N and it is also robust respect to noise contamination. Here we explore these very same problems for the case of the VG motif profile and we make a detailed comparison of its performance with the HVG motif profile in a range of situations.

A. Convergence properties for finite size series

In general, due to finite size effects, the estimated value of any feature fluctuates and deviates with respect to its asymptotic, expected value. For classical features such as the mean or the variance of a distribution, these deviations are bounded and vanish with series size with a speed quantified by the central limit theorem. The estimation of the motif frequencies can be quantitative effected by finite-size fluctuations and one can even observe missing motifs (motifs with estimated frequency $\Phi = 0$) which are not actually forbidden by the process but have not appeared by chance. This situation can be overemphasized in the presence of certain types of measurement noise.

Following an approach analogous to the one followed for the forbidden ordinal patterns in [27–29], we first perform a test to study the decay of missing motifs with the series size both in stochastic uncorrelated and correlated processes. In Figure 5 panel a) we plot $\langle R(N) \rangle$, the average number of missing motifs in a series of size N in the case of Gaussian white noise and colored (red) Gaussian noise (for the red noise we consider the AR(1) process with correlation length $r = 0.5$ discussed in section III). For both types of noise $\langle R(N) \rangle$ decays exponentially to zero and already with a series of about 80-100 data points we can exclude the possibility of detecting missing motifs (for both HVG and VG) due to finite size fluctuations even in the case of correlated noise.

As a second analysis, we explore the convergence speed of the estimated motif profile of uncorrelated and correlated stochastic series and of chaotic series (fully chaotic logistic map) of size N to the asymptotic profile solution given in section III. To do this we define the distance between the estimated 4-motif probabilities $\hat{\Phi}_m^4(N)$ and the asymptotic value $\Phi_m^4 = \lim_{N \rightarrow \infty} \hat{\Phi}_m^4(N)$. We use ℓ_1 norm and accordingly define

$$d(N) = \sum_m |\hat{\Phi}_m^4(N) - \Phi_m^4| \quad (24)$$

In Figure 5 panel b) we show the trend of $d(N)$ in log-log scale (results are averaged over 300 realizations). The average distance decreases like a power-law for all the processes considered, in agreement with a central-limit-theorem-like argument. For a series of $N = 10^3$ points $\langle d(N) \rangle$ is less than $5 \cdot 10^{-2}$ and the average distance $\langle d_m \rangle$ for each of the

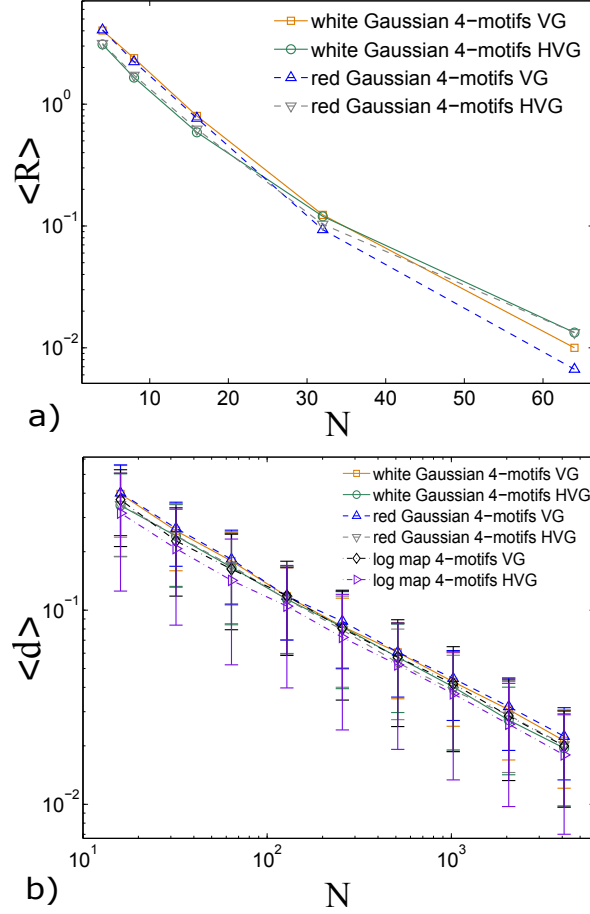


FIG. 5: Robustness of VG and HVG motif statistic respect to finite series size effects, in the case of Gaussian white noise and colored (red) Gaussian noise. Panel (a): Semi-log plot of the average number of missing motifs $\langle R \rangle$ vs series size N (each point is an average over 300 realizations of the corresponding process). $\langle R \rangle$ decays exponentially to zero, meaning that for $N \sim 100$ we can already exclude the possibility of detecting missing motifs due to finite size fluctuations for both type of noise. Panel (b): Log-log plot of the average distance $\langle d \rangle$ between the observed motif profile and the theoretical profile as a function of the series size N (results are averaged over 300 realizations). $\langle d \rangle$ decreases as a power-law for all the processes considered.

single components is less than 10^{-2} (not shown). These results suggest that VG and HVG motif profiles have very good convergence properties and are thus robust against finite size fluctuations.

B. Robustness against measurement noise

To test and compare the robustness of VG and HVG motif profiles when the effect of noise contamination combines with the finite size fluctuations we consider the fully chaotic logistic map dynamics x_t polluted with measurement (additive) noise η_t

$$\begin{cases} X_t = x_t + \eta_t \\ x_t = 4 \cdot x_{t-1} \cdot (1 - x_{t-1}) \\ \eta_t = r \cdot \eta_{t-1} + \sqrt{\alpha} \cdot \xi_t, \quad \xi_t \in \mathcal{N}(0, 1) \end{cases} \quad (25)$$

in the two cases where η_t is respectively white Gaussian noise ($r = 0$) or colored Gaussian noise ($r = 0.5$). For both cases $\alpha \in [0, 1]$ is the parameter which tunes the noise-to-signal ratio (NSR) of the process defined as

$$\text{NSR}(\alpha) = \frac{\sigma^2[\sqrt{\alpha} \cdot \xi]}{\sigma^2[x]} \quad (26)$$

where $\sigma^2[\sqrt{\alpha} \cdot \xi]$ and $\sigma^2[x]$ are respectively the theoretical variance of the white Gaussian noise $\sqrt{\alpha} \cdot \xi$ and the theoretical variance of the dynamics (signal) x (note that with this definition we are underestimating the NSR in the case of correlated noise where $\sigma^2[\eta] = \sigma^2[\sqrt{\alpha} \cdot \xi]/(1 - r^2)$). The robustness of the observed motif profile $\hat{\Phi}_i^4[X(N, \alpha)]$ for a single realization of the process with given N and α can be defined as the distance between this profile and the theoretical profile $\Phi_i^4[\eta(\alpha)]$ of the noise η for the given α .

$$\delta(N, \alpha) = \sum_m |\hat{\Phi}_m^4[X(N, \alpha)] - \Phi_m^4[\eta(\alpha)]|. \quad (27)$$

With such definition we expect $\delta(N, \alpha) \gg 0$ for low values of the NSR (dominant signal, $\hat{\Phi}_m^4[X(N, \alpha)] \simeq \Phi_m^4[x]$) and $\delta(N, \alpha) \simeq 0$ for high values of the NSR (dominant noise, $\hat{\Phi}_m^4[X(N, \alpha)] \simeq \Phi_m^4[\eta]$). Furthermore, $\delta(N, \alpha)$ is affected by finite size effects: if we assume to have few realizations N_r of the process X of small series size N , then we expect the variance $\sigma^2(\delta(N, \alpha))$ calculated over the realizations to be high. In particular we have to consider that a resolution limit δ^0 exists, such that when $\langle \delta(N, \alpha) \rangle_{N_r} = \delta^0$ we cannot say any more if the distance we measured is discriminating the signal x from the noise η or it is simply due to finite-size effects of the contamination noise η . We define this threshold δ^0 as the sum of the standard deviations of the estimated profile components $\hat{\Phi}_m^4[\eta]$ given N_r realizations of the noise process alone

$$\delta^0(N, \alpha) = \sum_m \sqrt{\langle (\hat{\Phi}_m^4[\eta(N, \alpha)] - \langle \hat{\Phi}_m^4[\eta(N, \alpha)] \rangle_{N_r})^2 \rangle_{N_r}} \quad (28)$$

It is thus convenient to work with the relative distance $\langle \delta \rangle_{N_r} / \delta^0$; when $\langle \delta \rangle_{N_r} / \delta^0 \leq 1$ we say that the resolution limit for the process $X(N, \alpha)$ -given its N_r realizations- is reached, and $\langle \delta \rangle_{N_r}$ is not any more a reliable indicator.

In Figure 6 panel a) we show the quantity $\langle \delta \rangle / \delta^0$ averaged over 300 realizations of the process X for different level of contamination NSR=[0, 0.2, 0.4, ..., 8] at fixed size $N=6400=100 \cdot 2^6$ (notice that for $N \geq 100$ we missing motifs are not found anymore) respectively for white Gaussian noise and colored Gaussian noise and for the HVG and the VG motif profile. The red solid line represents the resolution limit threshold for the process. We can see that the HVG and the VG motif profiles are more robust respect to noise contamination when this noise is correlated. In this situation the HVG motif profile seems to perform better than the VG motif profile, while in the case of uncorrelated Gaussian noise the VG profile seems in turn slightly more robust than the HVG profile.

The last step of this robustness analysis is to consider the usual situation where only very few realizations (often a single one) of the same process are available. Our aim is to define a useful indicator θ which estimates for any given value of the size N the maximum amount of noise contamination level for which a measure δ computed with only one realization of the process X can be considered somewhat reliable. We define this to be the value of the NSR such that $\langle \delta \rangle - \sigma(\delta) = \delta^0$, and thus

$$\theta(N) = \{\text{NSR}(\alpha) : \langle \delta(N, \alpha) \rangle - \sigma(\delta(N, \alpha)) = \delta^0\}. \quad (29)$$

$\theta(N)$ measures (in units of noise-to-signal ratio) the (statistical) reliability of the motif profile extracted from a single time series of size N of the signal x in the presence of measurement noise η .

In figure 6 (panel a) we plot $\theta(N = 6400)$ for white Gaussian noise in the case of VG by considering the $\langle \delta \rangle / \delta^0$ curve marked by orange squares and by taking the smallest value of NSR for which an orange error bar intersect the red line (the blue box highlights the region). We find $\theta \simeq 2.2$, meaning that when working with a single time series of the process X with size $N = 6400$, the $\langle \delta \rangle$ distance measured by using the VG motif profile is reliable up to a level of white Gaussian noise contamination α such that $\text{NSR}(\alpha) \simeq 2.2$. In Figure 6 (panel b) we report the estimated value of θ for the VG and HVG motif profiles in the case of white Gaussian noise and correlated Gaussian noise in function of the series size $N=100 \cdot 2, 100 \cdot 2^2, \dots, 100 \cdot 2^7$ (maximum noise contamination level considered was $\text{NSR}(\alpha)=8$). We can see that the motif profile is in general a robust measure respect to the combined effect of measurement noise and finite size: working with a single time series of only 3000 points of the process X we can extract both the VG and the HVG motif profiles and expect those features to be informative respect to the underlying chaotic signal x up to a level of measurement noise for which $\text{NSR}=1.5$ in the case of uncorrelated Gaussian noise and $\text{NSR}=3$ in the case of correlated Gaussian noise.

Also and as observed before (Figure 6b)), given the case of white Gaussian noise contamination the VG motif profile (orange squares) seems to perform slightly better than the HVG motif profile (green circles). For colored Gaussian noise the situation is the opposite and the HVG motif profile (reversed gray triangles) performs much better (almost a gap of one unit of NSR for $N > 1600$) than the VG motif profile (blue triangles). For both type of visibility graphs the motif profile is coherently more robust when polluted with colored noise than with white noise. This is probably due to the fact that white noise breaks up the correlation structure of the signal faster (respect to the size N) than

correlated noise. It is also interesting that both types of motif profiles are very sensible to the noise correlations although the different nature of the visibility algorithms.

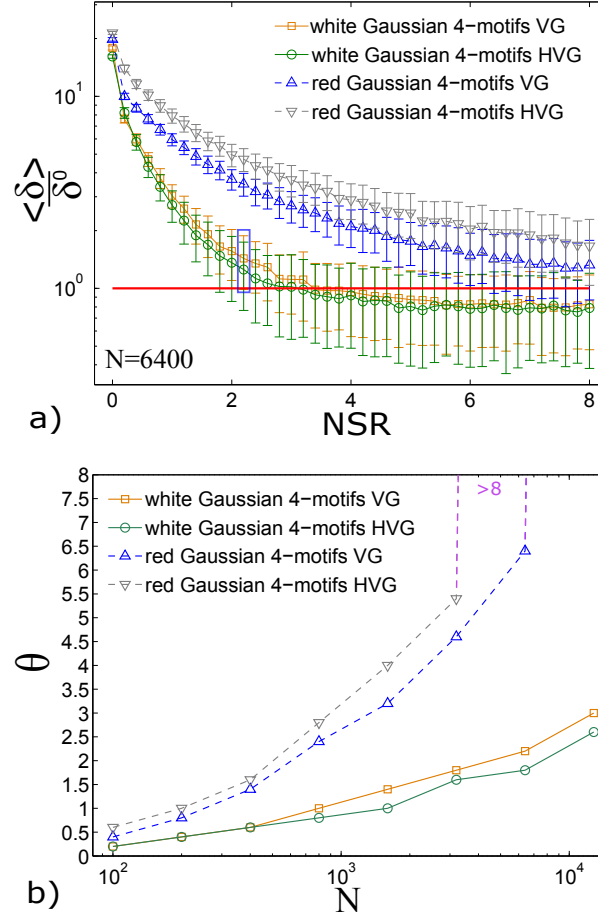


FIG. 6: Robustness of HVG and VG motif profiles respect to measurement noise. Panel a) the average distance between the motif profile extracted from the polluted chaotic dynamics X (see Eq.25) and the theoretical motif profile of the noise, normalized by the resolution limit threshold δ^0 for different level of contamination NSR (noise-to-signal ratio) at fixed size $N=6400$. When the curves reach the resolution limit (solid red line) we cannot consider any more the motif profile informative about the underlying chaotic dynamics due to the noise effects. Panel b) the estimated values of the measure θ in function of the time series size N indicating the maximum amount of NSR for which the motif profile extracted from a single realization of the process X is reliably informative respect to the chaotic signal. The HVG and the VG motif profiles are more robust respect to noise contamination when the noise is correlated (red Gaussian). In this situation the HVG motif profile seems to perform better than the VG motif profile, while in the case of uncorrelated Gaussian noise the VG profile seems in turn slightly more robust than the HVG profile.

V. CONCLUSIONS

Sequential visibility graph motifs are small subgraphs where nodes are in consecutive order within the Hamiltonian path that appear with characteristic frequencies for different types of dynamics. This concept was introduced recently [13] and a theory was developed to analytically compute the motif profiles in the case of horizontal visibility graphs (HVGs). In this work we have extended this theory to the realm of natural visibility graphs (VGs), a family of graphs where the previous amount of known exact results was practically null. We have been able to give a closed form for the 4-node VG motif profile associated to general one dimensional deterministic and stochastic processes with a smooth invariant measure or continuous marginal distribution, for the cases where the variables belong to a bounded or unbounded interval. In the case where the time series is empirical and one does not have access to the underlying dynamics, the methodology still provides a linear time ($O(N)$) algorithm to estimate numerically such profile. We have shown that the theory is accurate and that VG motifs have similar robustness properties as HVG, yet they

depend on the marginal distribution of the process and as such yield different profiles for different marginals. This is at odds with the results found for HVGs, where the motif profiles did not depend on the marginals as they behave as an order statistic.

The detection of such motifs from a visibility graph extracted from a time series can be seen as a process of dynamic symbolization of the series itself, where the alphabet of symbols is composed by different subgraphs (motifs) which encode information about both data relations and their temporal ordering in their link structure. The deep similarity between HVG motifs and the so called ordinal pattern analysis -which holds mainly due to the fact that HVG is an order statistic- vanishes for VG motifs, which therefore stand as a complementary tool for time series analysis, specially relevant when the marginals play a role in the analysis.

-
- [1] Zhang J, Small M, Complex network from pseudoperiodic time series: topology versus dynamics. *Phys. Rev. Lett.* **96**, 238701 (2006).
 - [2] Kyriakopoulos F, Thurner S, Directed network representations of discrete dynamical maps, in *Lecture Notes in Computer Science* **4488**, 625–632 (2007)
 - [3] Xu X, Zhang J, Small M Superfamily phenomena and motifs of networks induced from time series. *Proc. Natl. Acad. Sci. USA* **105**, 19601–19605 (2008).
 - [4] Donner R V, Zou Y, Donges J F, Marwan N, Kurths J Recurrence networks: a novel paradigm for nonlinear time series analysis. *New J. Phys.* **12**, 033025 (2010).
 - [5] Donner R V, et al. The Geometry of Chaotic Dynamics - A Complex Network Perspective. *Eur. Phys. J. B* **84**, 653–672 (2011).
 - [6] B. Bollobás, *Modern graph theory*, vol. 184 (Springer Science & Business Media, 2013).
 - [7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Physics reports **424**, 175 (2006).
 - [8] M. E. Newman, SIAM review **45**, 167 (2003).
 - [9] D. S. G. Pollock, R. C. Green, and T. Nguyen, *Handbook of time series analysis, signal processing, and dynamics* (Academic Press, 1999).
 - [10] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control* (John Wiley & Sons, 2015).
 - [11] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuno, Proceedings of the National Academy of Sciences **105**, 4972 (2008).
 - [12] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, Physical Review E **80**, 046103 (2009).
 - [13] J. Iacovacci, and L. Lacasa, Sequential visibility-graph motifs, *Phys. Rev. E* **93**, 4 (2016).
 - [14] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
 - [15] U. Alon, Nature Reviews Genetics **8**, 450 (2007).
 - [16] L. Lacasa, On the degree distribution of horizontal visibility graphs associated with Markov processes and dynamical systems: diagrammatic and variational approaches, *Nonlinearity* **27**, 9 (2014).
 - [17] A.M. Nunez, B. Luque, L. Lacasa, J.P. Gomez, A. Robledo, Horizontal Visibility graphs generated by type-I intermittency, *Phys. Rev. E*, **87** (2013) 052801
 - [18] A.M. Nunez, L. Lacasa, J.P. Gomez, Horizontal Visibility graphs generated by type-II intermittency, *J. Phys. A* **47**(3) (2013) 14
 - [19] B. Luque, L. Lacasa, F. Ballesteros, A. Robledo, Analytical properties of horizontal visibility graphs in the Feigenbaum scenario, *Chaos* **22**(1) (2012) 013109
 - [20] L. Lacasa, B. Luque, J. Luque, J.C. Nuño, The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion, *EPL (Europhysics Letters)* **86**, 3 (2009).
 - [21] L. Lacasa and R. Flanagan, Time reversibility of visibility from non-stationary processes, *Physical Review E* **92**, 02281 (2015).
 - [22] S. Severini, G. Gutin, T. Mansour, A characterization of horizontal visibility graphs and combinatorics on words, *Physica A* **390**(12) (2011) 2421–2428
 - [23] C. Bandt and B. Pompe, Physical review letters **88**, 174102 (2002).
 - [24] C. Bandt, Ecological modelling **182**, 229 (2005).
 - [25] J. Amigó, *Permutation complexity in dynamical systems: ordinal patterns, permutation entropy and all that* (Springer Science & Business Media, 2010).
 - [26] J. M. Amigó, L. Kocarev, and J. Szczepanski, Physics Letters A **355**, 27 (2006).
 - [27] J. M. Amigó, S. Zambrano, and M. A. Sanjuán, EPL (Europhysics Letters) **79**, 50001 (2007).
 - [28] L. C. Carpi, P. M. Saco, and O. Rosso, Physica A: Statistical Mechanics and its Applications **389**, 2020 (2010).
 - [29] O. A. Rosso, L. C. Carpi, P. M. Saco, M. G. Ravetti, A. Plastino, and H. A. Larrondo, Physica A: Statistical Mechanics and its Applications **391**, 42 (2012).